

# Indexing and Mapping of Proteins Using a Modified Nonlinear Sammon Projection

IZYDOR APOSTOL,<sup>1</sup> WOJCIECH SZPANKOWSKI<sup>2</sup>

<sup>1</sup> Baxter Hemoglobin Therapeutics Inc., Boulder, Colorado 80301

<sup>2</sup> Purdue University, Department of Computer Science, West Lafayette, Indiana 47907

Received 25 September 1998; accepted 26 February 1999

**ABSTRACT:** A modified Sammon algorithm was developed to display a relationship between proteins based on their amino acid composition. In the first stage of the method, a 19-dimensional compositional space of representative proteins was mapped into a two-dimensional space (2D) using the original Sammon projection creating a contour map. In the second stage, this contour map was used as a reference for new proteins projected into 2D. Data analysis showed that proteins belonging to the same structural classes formed characteristic and distinct clusters, which could be potentially useful in the prediction of protein structural classes. However, we observed significant overlapping of the clusters, which may explain the limited success of previous protein folding prediction based solely on amino acid composition. Regardless, the modified Sammon projections can generate a unique index for each individually projected protein related to its amino acid composition, which may be a useful tool in the exploratory classification of proteins. © 1999 John Wiley & Sons, Inc. *J Comput Chem* 20: 1049–1059, 1999

**Keywords:** nonlinear Sammon projection; amino acid composition; protein structural classes; indexing of proteins

## Introduction

The classification of proteins and the prediction of their structural classes are important tasks in the characterization of newly discovered proteins. Unfortunately, the classification of pro-

teins is limited only to a subset of proteins with enzymatic activities. Even for those, no general indexing has been widely accepted. Also, the prediction of protein folding classes based on primary sequence information remains a difficult task.<sup>1–9</sup>

The comparison of more than two protein sequences is generally not a straightforward process. Aligning them and measuring similarity/dissimilarity distances requires complicated computing in

Correspondence to: I. Apostol; e-mail: apostoi@baxter.com

multidimensional space equal to the length of the protein sequences. The comparison of proteins of different lengths creates an additional requirement of bringing them to the same multidimensional space, which necessitates the introduction of complicated sequence gaps. It is far simpler to compare the amino acid composition (AAC) of proteins. In this case, all proteins can be compared in the same 19D compositional space. An individual protein can be described as a point or a vector in compositional space determined by the molar fractions of the 20 amino acids.<sup>10</sup> By definition, the sum of the molar fractions of all components is equal to 1; therefore, only 19 components are independent. By leaving out one component, the protein can still be unambiguously determined in 19D space. Even in this reduced space, comparison of protein vectors is not easy, primarily due to the inability of humans to adequately visualize objects in spaces with greater than three dimensions. Several attempts have been made to use AAC information to predict protein folding classes.<sup>7,10-23</sup> In most cases, the compositional information has been brought into a linear format with the introduction of various weighting factors and averaging methods. These procedures resulted in a significant reduction of information that limited its predictive potential.

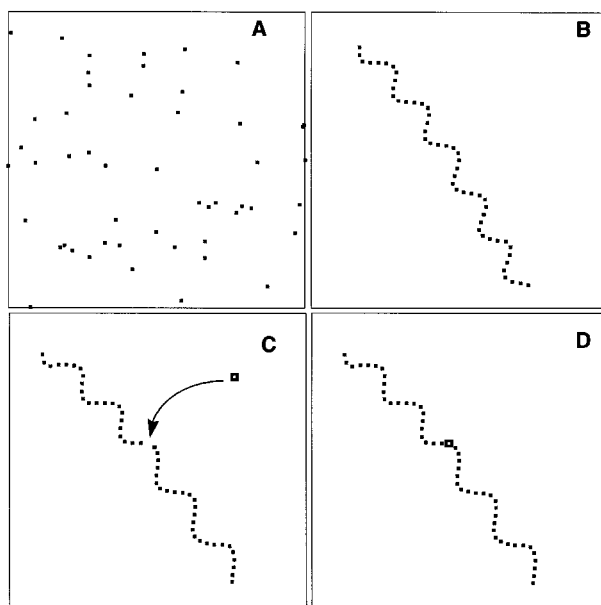
In this work we attempt to use Sammon mapping to project a 19D compositional space of proteins into a 2D space to observe the relationship between proteins within that newly created space. The algorithm works by projecting protein vectors from the compositional space onto a planar display such that the Euclidean distances between the projected images (points) approximates, as closely as possible, the corresponding distance in the original compositional space.<sup>24-26</sup> Neither averaging nor introduction of correction/weighting factors is necessary in this method. The ability of this algorithm to capture the essential features of protein sequence or structure similarities was recently demonstrated by Agrafiotis<sup>25</sup> and Barlow and Richards.<sup>26</sup>

## Results

### HELIX EXAMPLE OF SAMMON NONLINEAR PROJECTION

To explain the Sammon nonlinear projection algorithm, we start with a nonbiological example

to show how the algorithm preserves certain dependencies/shapes (e.g., helices) when projected from 3D to 2D space. Figure 1 displays the results obtained using the nonlinear Sammon algorithm to project 50 points distributed evenly along a three-dimensional helix. The parametric equations for this helix are:  $X = \cos Z$ ,  $Y = \sin Z$ ,  $Z = t/2^{1/2}$ . The points were distributed at one-unit intervals along the curve. To initiate the algorithm we selected 50 corresponding random points in 2D space (Fig. 1A). Each point was assigned to represent one of the 50 points on the helix (3D space). The application of the Sammon algorithm caused the 2D points to organize in such a fashion that the Euclidean distances between all pairs of points in 2D space closely reflect their Euclidean distances on the helix in 3D space. After 250 iterations for each point, using the steepest descent algorithm described in the Methods section ( $MF = 0.3$ ), the projected points formed a highly organized wave shape (Fig. 1B), as described by Sammon.<sup>24</sup> Figure 1C displays the results of an experiment in



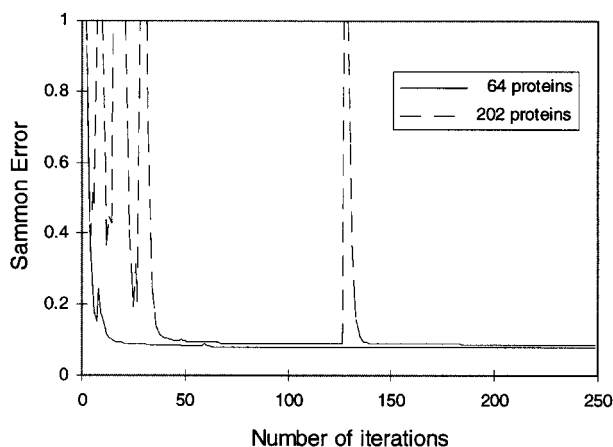
**FIGURE 1.** Sammon projection of a helix: (A) starting random distribution of 50 points; (B) optimized projection (map) of 50 points after 250 iterations; (C) contour map of 49 points (one selected point is missing from the map of 50 points) after 250 iterations; (D) projection of the missing 50th point into the contour map after 50 further iterations.

which one random point was excluded from the helix. The remaining 49 points were projected creating a gap in the projected wave shape. In the third experiment, the missing point was reintroduced. All 50 points were then projected but this time the  $X, Y$  coordinates of 49 points were fixed as in Figure 1C. That is, only the 50th point was optimized, greatly simplifying the calculations. After only 50 iterations, the missing point fell back into the gap and completed the wave shape. The orientation of the waves in Figure 1C and D are different from that in Figure 1B because the optimizations started from different randomly distributed points in 2D space, resulting in different projections.

These experiments demonstrate that a set of vectors in three-dimensional space can be projected into two-dimensional space, recreating the dependencies from a higher dimension even if one element of the set is missing. The missing element can then be added back to the precomputed set to complete the structure without the need to reoptimize the entire set *de novo*. This approach represents a significant advantage over traditional approaches because a full reoptimization "costs"  $n^2$  vs.  $n$  computations for one point optimization into the constant contour map. (We can now extrapolate this idea to compositional space of proteins.)

### THE SAMMON PROJECTION OF PROTEINS BELONGING TO DEFINITE STRUCTURE CLASSES

The AAC of 64 proteins classified by Chou<sup>13</sup> to specific structural classes were used to calculate the mol fractions of all 20 amino acids. Each protein was described as a unique vector in compositional space, with coordinate factors in the range of 0 to 1, which corresponds to the molar fraction for each particular amino acid. We numbered the 64 proteins as 1, 2, 3...64 according to the order listed in Tables 5–8 of Chou<sup>13</sup> or Tables 2–5 of Zhang and Chou.<sup>14</sup> All 64 protein vectors in the 19D composition space were then projected into a 2D space using the Sammon nonlinear algorithm. The error of the Sammon projection [Materials and Methods section, eq. (1)] plotted against the number of iterations using the steepest descent optimization ( $MF = 0.3$ ) is shown in Figure 2. A significant degree of optimization of this projection was seen after about 75 iterations, but occasionally a

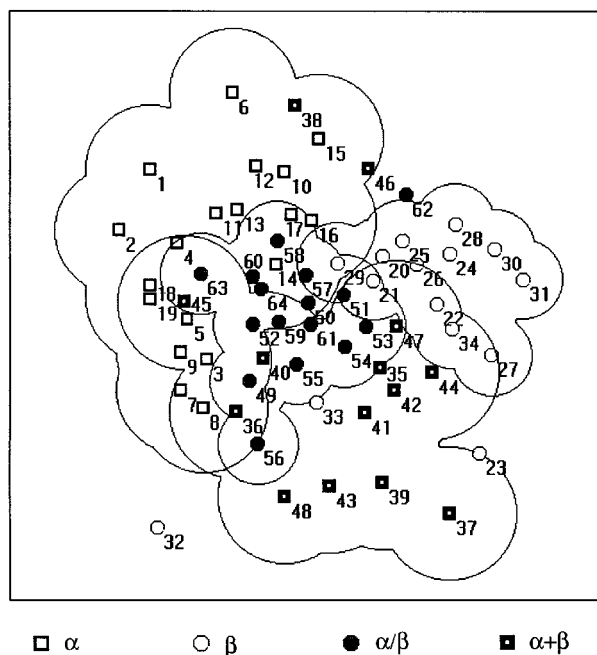


**FIGURE 2.** Error of Sammon projection ( $MF = 0.3$ ) plotted against number of iteration for sets of 64 (solid line) and 202 (dashed line) projected proteins.

burst of error was generated during later iterations. Typically, the results were analyzed following 250 iterations, well after the errors were resolved. Several projections of Chou's proteins have been made starting from different random distributions of points. In all cases the results were similar, primarily differing in the axial orientation of the projection and slightly in the resulting error. The best projection, obtained in these experiments based on the smallest Sammon's error (0.0746) is presented in Figure 3. The four different symbols correspond to four different structural classes: open squares (1–19)—19  $\alpha$  proteins; open circles (20–34)—15  $\beta$  proteins; filled squares (35–48)—14  $\alpha + \beta$  proteins; filled circles (49–64)—16  $\alpha/\beta$  proteins. It is apparent that the structural classes formed recognizable groupings. However, a few points are clear outliers, and are located outside the expected clusters. In general, points that were "misplaced" correspond to the proteins that were also misassigned by the Zhang and Chou prediction algorithm.<sup>14</sup> For example, point 32 (outside of any cluster) corresponds to Rubredoxin, which was classified as a  $\beta$  protein. Its unusual position can be explained by its low level of  $\beta$  structure (25%), lower than any other protein found in that set. The Zhang and Chou algorithm predicts  $\alpha/\beta$  structural class for Rubredoxin. Point 38, which represents a High-potential Iron Protein from the  $\alpha + \beta$  class, falls into the cluster of  $\alpha$  proteins in our projection. In this case, only 27% of the total pro-

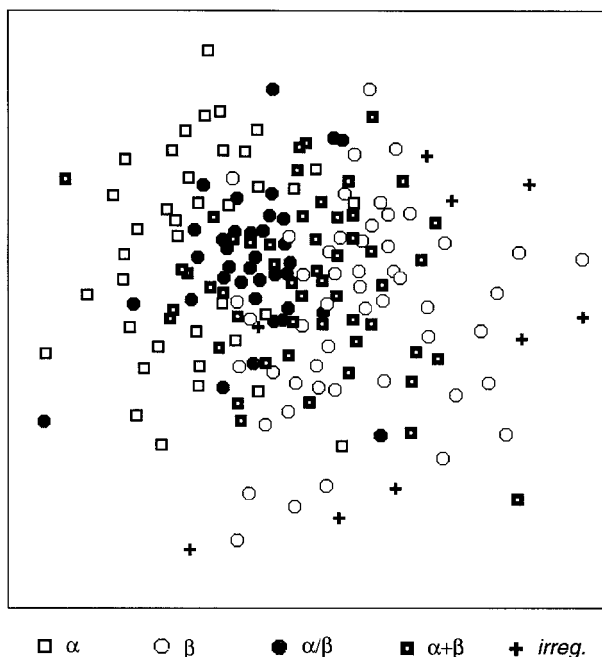
tein has a defined structure. Zhang and Chou placed this protein in the  $\alpha$  structural class. Similar arguments could be made for other outlying points: 23, 33, 46, and 62. This may suggest that the presence of irregular protein structure can significantly influence the AAC, and may result in inaccurate class assignment/prediction.

The relationship within each set of proteins was analyzed by complete-linkage cluster analysis. Cluster analysis can provide an objective automated way to group objects into clusters in multi-dimensional spaces. The correlation coefficient of each cluster corresponds to the maximum distance in which two objects are still considered to have similar properties. We defined the correlation coefficient as the median 2D Euclidean distance for the set after excluding outliers (points: 23, 32, 33, 38, 46, and 62). The results of the cluster analysis are presented as outlines surrounding each set in Figure 3.



**FIGURE 3.** 2D contour map of compositional space of Chou's 64 proteins<sup>13</sup> belonging to four different folding classes. Open squares—19  $\alpha$  proteins; open circles—15  $\beta$  proteins; filled squares—14  $\alpha + \beta$  proteins; filled circles—16  $\alpha / \beta$  proteins. Outlines were drawn based on cluster analysis for each structural class. Optimization was terminated after 250 iterations (MF = 0.3). The error of the projection was 0.0746.

In the process of data analysis we realized that the set of 64 Chou's proteins could be too small to make a generalization about all proteins. Therefore, we examined another set of 193 proteins that was used by Chou<sup>19</sup> in his latest predictions of structural classes utilizing Mahalanobis distances. This includes 129 proteins from a learning set and 64 proteins from a testing set. Proteins in both sets were selected using more precise constraints of classification.<sup>19</sup> Compositional space of all 193 proteins was then projected into 2D space using the nominal Sammon nonlinear algorithm. The projection was terminated after 500 iterations (MF = 0.3), and the result is shown in Figure 4. We used the same symbols as in Figure 3 to designate folding classes: open squares—39  $\alpha$  proteins; open circles—52  $\beta$  proteins; filled squares—54  $\alpha + \beta$  proteins; filled circles—39  $\alpha / \beta$  proteins; crosses—9  $\zeta$  (irregular) proteins. This result was not significantly different from the previously described pro-



**FIGURE 4.** 2D contour map of compositional space of Chou's 193 proteins<sup>19</sup> belonging to different folding classes: open squares—39  $\alpha$  proteins; open circles—52  $\beta$  proteins; filled squares—54  $\alpha + \beta$  proteins; filled circles—39  $\alpha / \beta$  proteins; crosses—9  $\zeta$  (irregular) proteins. Optimization was terminated after 500 iterations (MF = 0.3). The error of the projection was 0.151.

jection of 64 proteins. As seen previously, the different structural classes formed distinct clusters in 2D space. The  $\alpha$  and  $\beta$  clusters were well separated. The  $\alpha/\beta$  and  $\alpha + \beta$  proteins formed smooth transitions between the  $\alpha$  and  $\beta$  clusters resulting in partial overlapping of structural clusters. Irregular proteins were projected outside of the  $\beta$  class. As before, several points were clear outliers and were not located within their expected clusters.

Very similar results were obtained for a different set of 202 proteins. These proteins were arbitrarily selected from SCOP data base available on the World Wide Web (<http://www.pdb.bnl.gov.scop>).<sup>27</sup> Again, the four structural classes formed distinct but overlapping clusters in 2D space (data not shown).

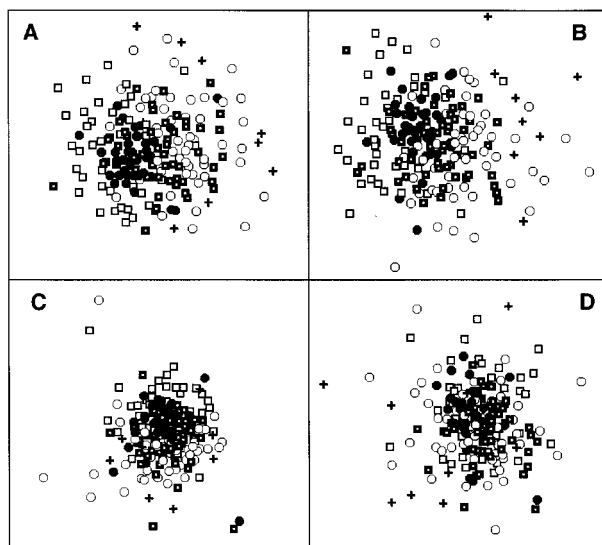
These results support the early hypothesis that the folding of proteins is determined by their AAC.<sup>7,10-17,19,21</sup> However, we observed a significant degree of overlapping between clusters (especially for  $\beta$  and  $\alpha + \beta$ ), which may explain the inaccuracy of the previous protein folding prediction based solely on AAC. The irregular portions found in all proteins may contribute significantly to the overall ambiguity.

#### THE SAMMON PROJECTION USING MODIFIED DISTANCES OR A REDUCED ALPHABET OF AMINO ACIDS

In hope of improving clustering of structural classes, we explored several different ways of defining distance between proteins in compositional space. We examine Minkowski's  $l^r$  distance

$$D_{pq}^* = \left[ \sum_{i=1}^{20} (P_i - Q_i)^r \right]^{1/r}$$

for  $r$  ranging from 0.5 to 64 (see Materials and Methods). Actually, Euclidean distance used in all the other experiments is a special case of Minkowski's definition with  $r = 2$ . Figure 5A and B show the examples of the Sammon projection for  $r$  equal to 1 and 16, respectively. No significant differences were found in the distribution or separation of structural classes. In fact, with increased value of  $r$  the projected map resulted in greater overlapping of structural clusters. These data indicate that the application of Minkowski's general distance definitions does not offer any advantage over the intuitive Euclidean distance definition.



**FIGURE 5.** Sammon projection of Chou's 193 proteins<sup>19</sup> after 250 iterations using (MF = 0.3): (A) Minkowski distance for  $r = 1$ ; (B) Minkowski distance for  $r = 16$ ; (C) Euclidean distance modified by weighting factors;<sup>17</sup> (D) a reduced amino acid alphabet.<sup>18</sup> The symbols corresponding to folding classes are the same as in Figure 4.

Additionally, we examined the modification of Euclidean distance by using weighting factors introduced by Chun-Ting et al.<sup>17</sup> In this particular case the calculation of Euclidean distance was modified as follows:

$$D_{pq}^* = \sqrt{\sum_{i=1}^{20} w_i (P_i - Q_i)^2}$$

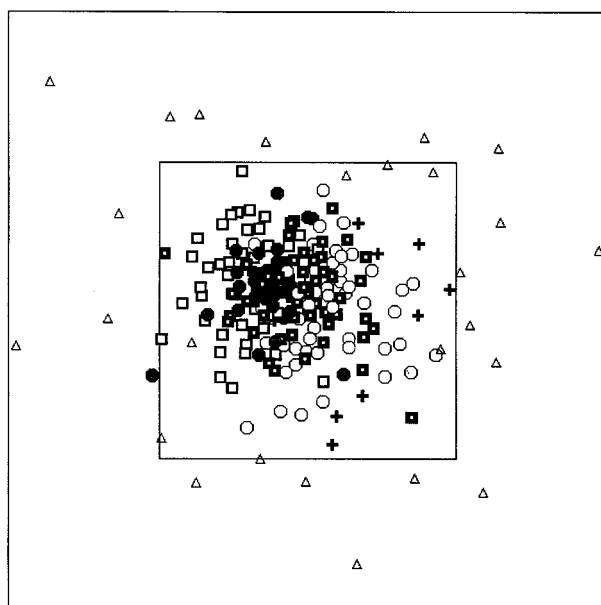
The distances in compositional space were calculated as the product of  $w_i$ , the weighting factor, and the Euclidean distance for each amino acid. Chun-Ting et al.<sup>17</sup> performed predictions of the structural class of proteins from AAC based on a linear programming approach using these weighting factors. We sought to determine whether weighting factors would affect nonlinear projections. The results (Fig. 5C) indicated that these weighting factors did not improve clustering of the proteins belonging to the same structural class.

Landes and Risler reported the successful use of a reduced amino acid alphabet in searching protein data bases.<sup>18</sup> They reduced the 20 amino acids

alphabet to 10 symbols: (A = T = S), C, (D = E = N), (F = Y), G, H, (I = L = M = V), (K = Q = R), P, W. We were interested to see if the reduction of the original compositional space into a nine-dimensional space would affect the clustering of the projected proteins. The application of the reduced alphabet into the Sammon projection of 193 proteins resulted in an irregular distribution of points (Fig. 5D). Clusters were still evident, but were not as clearly separated as in the case of the nominal projection from 19D compositional space (Fig. 4).

### PROJECTION OF PROTEINS WITH UNUSUAL AMINO ACID COMPOSITION

We investigated whether Sammon's projection could be used to predict the folding class of "unknown" proteins. Several hundred random proteins from the Pir1 protein database were projected individually into the developed map shown in Figure 4. In this modified projection method, the coordinates of the 193 representative proteins that formed the contour map were held constant, while the coordinates of new proteins were optimized by the Sammon algorithm. Each newly projected protein was treated as the 194th element. Several projected proteins were found outside the area occupied by the four clusters of the representative set. Most of these proteins appeared to be small and/or showed unusual amino acid composition. Several of these proteins were described as unusual by Cornish-Bowden in his work on dependencies between the size and AAC of proteins.<sup>28</sup> Additionally, we observed that if each individual protein was projected several times, the results of the optimizations were significantly different. We believe that this was due to a lack of reference points outside the area occupied by the set of 193 representative proteins. To overcome that ambiguity, a subset of 27 extra proteins with unusual AAC were selected from the Pir1 protein data base and added to the representative set. These proteins or peptides were arbitrarily selected if their length exceeded 10 amino acids and their X,Y indices fell outside the original clusters. Adding unusual proteins expanded the representative set from 193 to 220. The expanded set was then projected using the original Sammon algorithm, forming a new expanded contour map (Fig. 6). More than 2000 proteins of randomly selected from Pir1 was then



□  $\alpha$     ○  $\beta$     ●  $\alpha/\beta$     ■  $\alpha+\beta$     + *irreg.*    △ *unusual*

**FIGURE 6.** 2D contour map of 220 proteins. The set of 193 proteins (Fig. 3) was appended with additional 27 proteins with unusual composition (Appendix A). After 500 iteration (MF = 0.3) the error of the Sammon projection was 0.135.

projected into the new extended contour map using the modified Sammon mapping procedure. This time we observed significantly less variability from multiple projections of the same proteins. Each protein gave a distinct point in the 2D map. Thus, it appeared that each newly projected protein could be characterized by unique X,Y coordinates on a 2D map, reflecting its unique amino acid composition. This method could offer a new way to classify proteins based solely on AAC.

### PROJECTION OF AMINO ACID COMPOSITION OF HEMOGLOBINS INTO THE EXTENDED CONTOUR MAP

The amino acid composition of porcine hemoglobin alpha chain was projected into the extended contour map as an additional 221th protein. As previously described, the X,Y coordinates of the 220 proteins from the extended contour map were held as fixed points. The projected porcine hemoglobin fell into the cluster of the  $\alpha$  proteins,

close to the other hemoglobins used in the representative set. The observed index for the  $\alpha$  chain of the porcine hemoglobin was 0.390:0.630, which is very close to the other hemoglobins used in the representative set. This operation of the projection of the 221th protein was repeated for the  $\beta$  chain of porcine hemoglobin and 320 other different  $\alpha$  and  $\beta$  globins with similar results. All of the projected  $\alpha$  and  $\beta$  chains formed a distinct well-defined group inside the cluster of  $\alpha$  proteins as expected for hemoglobins (Fig. 7). Similarly, 47 protein-tyrosine kinases and 14 serine proteinases were projected into the extended contour map. These also formed distinct groups inside the  $\alpha/\beta$  and  $\beta$  classes respectively, as would be predicted for those protein families (Fig. 7).

Additionally, a subset of 60 proteins used by Nakashima et al.<sup>10</sup> for their folding class predictions were projected into the extended contour map (data not shown). Interestingly, almost all these proteins fell into the clusters to which they

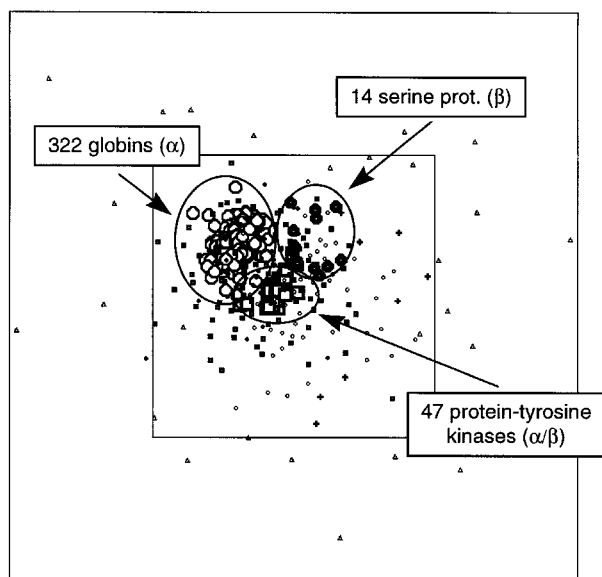
had been previously classified. Together, our results indicate that the Sammon nonlinear projection can be used to predict the structural classes of unknown proteins, although some ambiguity in the assignment remains due to the overlapping of structural class clusters.

## Conclusion

The nonlinear Sammon mapping algorithm may be a useful tool to examine the complicated multi-dimensional systems in protein science. When applied to project 19D compositional space of proteins, it can provide a new way of mapping and indexing. We have demonstrated that proteins with similar functionality can be mapped to the same region in 2D space. This method may have application in the prediction of folding classes and potentially functional properties of newly sequenced proteins based on compositional indices.

Our results suggest that the prediction of protein folding based on amino acid composition may never overcome certain limitations, such as overlapping clusters. Although different structural classes of proteins form distinct clusters when projected into 2D space, these clusters have a tendency to overlap. This overlapping of clusters may result in the ambiguous assignment of structural classes of unknown proteins, especially proteins sharing significant contributions of irregular fold. The validity of this approach very strongly depends on whether the examined set of proteins has a broad enough distribution of AAC. Therefore, it is possible that further expansion of the representative/learning set with detailed verification of the protein folds may limit overlapping and improve prediction accuracy.

We would like to point out the computational advantage of the contour map. The original Sammon algorithm,<sup>24</sup> and its application to the protein classification as proposed by Agrafiotis,<sup>25</sup> required  $N^2$  computations per iteration of the steepest descent algorithm (as discussed in the Materials and Methods section), where  $N$  is the number of projected proteins. In our modified method, representative proteins have been projected onto a precomputed contour map (in our experiments  $N = 220$ ), and new proteins were added by comparing them only individually to these  $N$  points. Therefore, the



**FIGURE 7.** Projection of: 322 different alpha and beta hemoglobin subunits (open circles), 47 protein-tyrosine kinases (open squares), and 14 serine proteinases (filled circles) into the 220 proteins contour map. Fifty rounds of optimization (MF = 0.3) was performed for each added protein.

projection of compositional space for a new protein costs only "N" computations per iteration. This difference translates to a significant saving in computation time, especially for large sets of proteins.

## Materials and Methods

### CALCULATION OF DISTANCES BETWEEN PROTEINS IN THE COMPOSITION SPACE

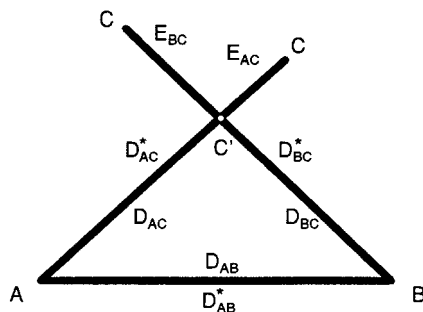
In the first experiment presented in Figure 1 we used the standard Euclidean distance between two points in 3D space. The distance between two proteins (level of dissimilarity) was computed based on the amino acid composition, according to the Minkowski  $l^r$  distance<sup>21</sup> in eq. (1):

$$D_{pq}^* = \left[ \sum_{i=1}^{20} (P_i - Q_i)^r \right]^{1/r} \quad (1)$$

where  $P_i$  and  $Q_i$  are the mol fraction of the  $i$ th kind of amino acid in the proteins  $P$  and  $Q$ , respectively. Each protein corresponds to a point with coordinates given by the mol fraction of the 20 constituent amino acids. Except where noted, the Minkowski distance was simplified to Euclidean distance with  $r = 2$ .

### THE SAMMON PROJECTION

The projection of 19D compositional space onto the 2D Euclidean space was obtained according to the original Sammon nonlinear projection algorithm.<sup>24-26</sup> This algorithm attempts to approximate in the best possible way (i.e., in the square error sense) a relationship between points in a multidimensional space when projected into 2D spaces. A detail description of the algorithm has been presented by Sammon,<sup>24</sup> so we only illustrate its meaning using a simple example. As shown below, the distance relationship in a higher dimension cannot be fully preserved in a lower dimensional space. For example, imagine three points A, B, and C in 3D space with given distances between them, defined as  $D_{AB}^*$ ,  $D_{AC}^*$ , and  $D_{BC}^*$ . These distances cannot be preserved in the projected 2D space ( $D_{AB}$ ,  $D_{AC}$ , and  $D_{BC}$ ). The best you can do is to preserve the distance between points A and B and minimize the error ( $E_{AC}$  and  $E_{BC}$ ) when locating point C.



In general, the goal of Sammon's algorithm is to minimize the discrepancy in "distance," which was defined as the error of the projection:

$$E(x, y) = \frac{1}{c} \sum_{i < j}^n \frac{(D_{ij}^* - D_{ij}(x, y))^b}{D_{ij}^*} \quad (2)$$

where  $D_{ij}^*$  is a distance in original space (19D compositional space) computed according to eq. (1),

$$D_{ij}(x, y) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

is the Euclidean distance in 2D space,  $b$  is a parameter,  $c$  is constant, and  $n$  is the number of proteins in the representative set. It should be pointed out that the parameter  $b$  can model a variety of situations.<sup>29</sup> Throughout the computation, as in Agrafiotis,<sup>25</sup> we have assumed that  $b = 2$  and

$$c = \sum_{i < j}^n D_{ij}^*.$$

To find optimal coordinates of a point  $(x, y)$ , a numerical optimization procedure called the steepest descent algorithm was applied. In this method, the optimal solution was found in several iteration starting from a (random) initial point. In each iteration we moved towards the gradient of the function  $E(x, y)$ . In the  $m$ th iteration we computed the  $m$ th estimate of  $E(x, y)$ , written as  $E_{(x, y)}^{(m)}$ . In the next iteration new coordinates of each point  $x^{(m+1)}$ ,  $y^{(m+1)}$  were computed according to



the following formula:

**for**  $m = 1$  **to** number of iterations

**for**  $i = 1$  **to**  $n$

$$x_i^{(m+1)} = x_i^{(m)} - \text{MF} \frac{\left| \frac{\partial E_{(x,y)}^{(m)}}{\partial x} \right|}{\left| \frac{\partial^2 E_{(x,y)}^{(m)}}{\partial x^2} \right|}$$

$$y_i^{(m+1)} = y_i^{(m)} - \text{MF} \frac{\left| \frac{\partial E_{(x,y)}^{(m)}}{\partial y} \right|}{\left| \frac{\partial^2 E_{(x,y)}^{(m)}}{\partial y^2} \right|}$$

**end**

**end.**

where MF ("magic factor") is an experimentally determined coefficient. The first and the second derivative of  $E(x, y)$  with respect to  $x$  are shown below (in a similar manner, one can compute the derivatives with respect to  $y$ ):

$$\frac{\partial E_{(x,y)}^{(m)}}{\partial x} = \frac{-2}{c} \sum_{\substack{i=1 \\ i \neq j}}^n \frac{(D_{ij}^* - D_{ij}(x, y))}{D_{ij}^* D_{ij}(x, y)} (x_i - x_j) \quad (3)$$

$$\frac{\partial^2 E_{(x,y)}^{(m)}}{\partial x^2} = \frac{-2}{c} \sum_{\substack{i=1 \\ i \neq j}}^n \frac{1}{D_{ij}^* D_{ij}(x, y)} \times \left[ (D_{ij}^* - D_{ij}(x, y)) - \frac{(x_i - x_j)^2}{D_{ij}(x, y)} \left( 1 + \frac{D_{ij}^* - D_{ij}(x, y)}{D_{ij}(x, y)} \right) \right] \quad (4)$$

In implementation of the steepest descent method we stopped the iteration procedure after 250–500 cycles.

In the experiments, when additional protein ( $n + 1$ ) was projected into the contour map, the same algorithms were used, except that the iteration procedure was used only to optimize the distance of the new protein without affecting the distances already optimized between elements of the learning set. The following modified formula was used:

**for**  $m = 1$  **to** number of iterations

$$x_{n+1}^{(m+1)} = x_{n+1}^{(m)} - \text{MF} \frac{\left| \frac{\partial E_{(x,y)}^{(m)}}{\partial x} \right|}{\left| \frac{\partial^2 E_{(x,y)}^{(m)}}{\partial x^2} \right|}$$

$$y_{n+1}^{(m+1)} = y_{n+1}^{(m)} - \text{MF} \frac{\left| \frac{\partial E_{(x,y)}^{(m)}}{\partial y} \right|}{\left| \frac{\partial^2 E_{(x,y)}^{(m)}}{\partial y^2} \right|}$$

**end.**

In addition, the summation of error  $E_{(x,y)}$  [eq. (1)] is over the single index  $i$  (this is only in terms of the sum).

## COMPUTER PROGRAMS AND DATA FILES

The source code of the computer programs were written in Borland Turbo Pasca<sup>®</sup> version 7. Executable Windows<sup>®</sup> versions of programs (Helix.exe and SammProj.exe) used in this article and corresponding contour maps, lists of proteins, and their indices are available at <http://www.cs.purdue.edu/people/spa/>.

## Acknowledgments

We express our special thanks to Marcin Apostol for many insightful suggestions during the writing of source codes for computer programs. We also thank our colleagues Dr. Spencer Anthony-Cahill, Dr. Jeff Etter, Dr. Greg Flynn, Dr. Doug Lemon, Dr. Michael Weickert, and the anonymous reviewers for critical reviews and helpful discussions.

## Appendix A

Amino acid composition of the 27 unusual proteins (unusual amino acid composition). The data for each protein contain two lines: the first line indicates its length, Pir code and name. The second line gives the frequencis of 20 amino acids according to the alphabetical order of the single amino acid code: ACDEFGHIKLMNPQRSTVWY.

430	PI;HHBYD8	Heat shock protein DDR48 - Yeast ( <i>Saccharomyces cerevisiae</i> )	0.70	1.16	11.40	0.00	0.70	6.98	0.23	0.70	23.26	0.00	1.63	1.40	26.51	0.47	0.93	0.00	10.00;	
0.47	0.00	13.49	Prothymosin alpha - Human	0.00	0.92	7.34	0.92	0.00	5.50	0.92	1.83	1.83	3.67	5.50	4.59	0.00	0.00;			
109	PI;TNNHUA	Prothymosin alpha - Human	0.00	0.92	7.34	0.92	0.00	5.50	0.92	1.83	1.83	3.67	5.50	4.59	0.00	0.00;				
10.09	0.00	17.43	Proteinase inhibitor (Bowman-Birk) D-II - soybean	0.00	1.33	1.33	1.33	5.33	4.00	2.67	6.67	5.33	6.67	16.00	4.00	0.00	4.00;			
75	PI;TISYD2	Proteinase inhibitor (Bowman-Birk) D-II - soybean	0.00	1.33	1.33	1.33	5.33	4.00	2.67	6.67	5.33	6.67	16.00	4.00	0.00	4.00;				
0.00	18.67	14.67	PI;FECFFerredoxin - <i>Chlorobium</i> sp.	0.00	1.64	8.20	0.00	1.64	0.00	1.64	0.00	3.28	40.98	14.75	1.64	6.56	0.00	4.92;		
61	PI;FECFFerredoxin - <i>Chlorobium</i> sp.	0.00	1.64	8.20	0.00	8.20	0.00	1.64	0.00	1.64	0.00	3.28	40.98	14.75	1.64	6.56	0.00	4.92;		
19.67	14.75	3.28	11.48	1.64	8.20	0.00	8.20	0.00	1.64	0.00	1.64	0.00	3.28	40.98	14.75	1.64	6.56	0.00	4.92;	
61	PI;PIHUOFF	Basic proline-rich peptide P-F - Human	0.00	0.00	22.95	0.00	0.00	8.20	0.00	0.00	3.28	40.98	14.75	1.64	6.56	0.00	0.00	0.00;		
1.64	0.00	0.00	0.00	0.00	22.95	0.00	0.00	8.20	0.00	0.00	3.28	40.98	14.75	1.64	6.56	0.00	0.00	0.00;		
51	PI;HSEBOS	Sperm histone - Bovine	0.00	1.96	3.92	1.96	1.96	0.00	1.96	0.00	1.96	0.00	1.96	50.98	3.92	5.88	3.92	0.00	3.92;	
1.96	13.73	0.00	0.00	1.96	3.92	1.96	1.96	0.00	1.96	0.00	1.96	0.00	1.96	50.98	3.92	5.88	3.92	0.00	3.92;	
48	PI;EWBY8	H+-transporting ATP synthase	0.00	0.00	0.00	0.00	8.33	2.08	25.00	8.33	2.08	6.25	6.25	4.17	6.25	4.17	0.00	6.25;		
0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.33	2.08	25.00	8.33	2.08	6.25	6.25	4.17	6.25	4.17	0.00	6.25;		
38	PI;C32038	mu-agatoxin III - funnel-weaving spider ( <i>Agelenopsis aperta</i> )	0.00	0.00	10.53	0.00	0.00	0.00	0.00	2.63	0.00	5.26	2.63	10.53	13.16	0.00	2.63	2.63	10.53;	
7.89	21.05	10.53	0.00	0.00	10.53	0.00	0.00	0.00	0.00	2.63	0.00	5.26	2.63	10.53	13.16	0.00	2.63	2.63	10.53;	
14	PI;QVHHMM	mastoparan M - hornet ( <i>Vespa mandarinia</i> )	0.00	0.00	0.00	0.00	14.29	21.43	28.57	0.00	7.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00;	
28.57	0.00	0.00	0.00	0.00	0.00	0.00	14.29	21.43	28.57	0.00	7.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00;	
13	PI;JUTG3	Tremorgen a-13 - <i>Basidiomycete</i> ( <i>Tremella mesenterica</i> )	7.69	0.00	38.46	0.00	0.00	0.00	0.00	0.00	7.69	7.69	0.00	7.69	7.69	0.00	7.69	0.00	0.00;	
0.00	7.69	7.69	0.00	0.00	38.46	0.00	0.00	0.00	0.00	0.00	7.69	7.69	0.00	7.69	7.69	0.00	7.69	0.00	0.00;	
11	PI;XASNBA	Bradykinin-potentiating peptide B - <i>Mamushi</i>	0.00	0.00	0.00	0.00	9.09	9.09	9.09	0.00	0.00	45.45	9.09	9.09	0.00	0.00	0.00	0.00	0.00;	
0.00	0.00	0.00	0.00	0.00	0.00	0.00	9.09	9.09	9.09	0.00	0.00	45.45	9.09	9.09	0.00	0.00	0.00	0.00	0.00;	
10	PI;AKLQ	Adipokinetic hormone - Migratory locust	0.00	0.00	10.00	0.00	0.00	0.00	0.00	0.00	20.00	10.00	10.00	0.00	0.00	20.00	0.00	10.00	0.00;	
0.00	0.00	0.00	0.00	0.00	10.00	0.00	0.00	0.00	0.00	0.00	20.00	10.00	10.00	0.00	0.00	20.00	0.00	10.00	0.00;	
10	PI;SPGPNK	neuromedin K - pig	0.00	0.00	10.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00;	
0.00	0.00	20.00	0.00	0.00	10.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00;	
259	PI;TPRPTS	Troponin T, skeletal muscle - Rabbit	0.00	0.00	0.00	0.00	14.29	21.43	28.57	0.00	7.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00;	
10.04	0.00	5.79	18.15	1.93	3.09	2.32	3.09	15.06	7.34	1.93	1.93	3.47	3.86	9.65	3.47	2.32	4.25	0.77	1.54;	
61	PI;SMHUIF	Metallothionein 1F - Human	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.64	1.64	3.28	1.64	0.00	14.75	3.28	4.92	0.00	0.00;
8.20	32.79	3.28	3.28	0.00	8.20	0.00	0.00	0.00	13.11	0.00	1.64	1.64	3.28	1.64	0.00	14.75	3.28	4.92	0.00	0.00;
13	PI;UNBONe	neurotensin - bovine	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00;	
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00;	
14	PI;QWMAVV	mastoparan - yellowjacket ( <i>Vespula lewisii</i> )	0.00	0.00	0.00	0.00	0.00	14.29	21.43	28.57	0.00	7.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00;	
28.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	14.29	21.43	28.57	0.00	7.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00;	
22	PI;MXKN1	mu-conotoxin GIIIA - cone shell ( <i>Conus geographus</i> )	0.00	0.00	0.00	0.00	0.00	0.00	18.18	0.00	0.00	0.00	13.64	9.09	13.64	0.00	4.55	0.00	0.00;	
4.55	27.27	9.09	0.00	0.00	0.00	0.00	0.00	0.00	18.18	0.00	0.00	0.00	13.64	9.09	13.64	0.00	4.55	0.00	0.00;	
82	PI;QPBmicro	glutamic acid-rich protein - bovine	0.00	0.00	0.00	0.00	0.00	0.00	12.20	0.00	0.00	0.00	2.44	2.44	0.00	1.22	2.44	0.00	0.00;	
18.29	0.00	4.88	47.56	0.00	8.54	0.00	0.00	12.20	0.00	0.00	0.00	2.44	2.44	0.00	1.22	2.44	0.00	0.00	0.00;	
40	PI;FDFI8	Antifreeze protein GS-8 - Grubby sculpin	0.00	0.00	2.50	0.00	2.50	0.00	2.50	5.00	0.00	2.50	0.00	2.50	0.00	5.00	0.00	0.00	0.00;	
60.00	0.00	5.00	2.50	0.00	2.50	0.00	2.50	0.00	2.50	5.00	0.00	2.50	0.00	2.50	0.00	5.00	0.00	0.00	0.00;	
291	PI;EEWTC	gamma-gliadin B precursor - wheat	1.03	4.47	2.75	1.37	6.87	1.37	8.25	2.06	1.37	15.12	30.24	1.37	6.19	3.09	5.15	0.69	1.03;	
4.12	2.75	0.69	1.03	4.47	2.75	1.37	6.87	1.37	8.25	2.06	1.37	15.12	30.24	1.37	6.19	3.09	5.15	0.69	1.03;	
29	PI;SNUMP	sillucinin - <i>Rhizomucor pusillus</i>	0.00	0.00	13.79	0.00	3.45	6.90	3.45	0.00	3.45	3.45	3.45	10.34	3.45	3.45	3.45	6.90;		
3.45	27.59	0.00	0.00	0.00	13.79	0.00	3.45	6.90	3.45	0.00	3.45	3.45	3.45	10.34	3.45	3.45	3.45	6.90;		
221	PI;KGZQH	histidine/alanine-rich protein - <i>Plasmodium falciparum</i>	0.90	3.17	3.62	27.15	0.45	2.71	4.98	0.45	12.67	0.00	0.45	4.52	1.36	1.81	0.00	0.00;		
27.60	0.45	7.24	0.90	3.17	3.62	27.15	0.45	2.71	4.98	0.45	12.67	0.00	0.45	4.52	1.36	1.81	0.00	0.00;		
61	PI;DNVPB	DNA-binding protein - budgerigar fledgling disease virus	0.00	0.00	1.64	1.64	0.00	1.64	34.43	1.64	3.28	14.75	6.56	14.75	6.56	3.28	0.00	0.00	0.00;	
9.84	0.00	0.00	0.00	0.00	1.64	1.64	0.00	1.64	34.43	1.64	3.28	14.75	6.56	14.75	6.56	3.28	0.00	0.00	0.00;	
44	PI;W5WLE	E5 protein - bovine papillomavirus type 1	2.27	15.91	4.55	2.27	0.00	0.00	34.09	4.55	2.27	4.55	2.27	0.00	2.27	4.55	4.55	2.27;		
4.55	4.55	2.27	2.27	15.91	4.55	2.27	0.00	0.00	34.09	4.55	2.27	4.55	2.27	0.00	2.27	4.55	4.55	2.27;		
660	PI;QQBE3	BHFI protein - human herpesvirus 4 (strain B95-8)	2.27	0.15	15.76	2.42	0.00	0.00	2.88	0.15	1.06	22.12	5.61	13.94	5.30	1.06	1.06	0.15;		
16.06	2.58	2.12	2.27	0.15	15.76	2.42	0.00	0.00	2.88	0.15	1.06	22.12	5.61	13.94	5.30	1.06	1.06	0.15;		
65	>PI;VHNVBM	nucleocapsid protein - <i>Bombyx mori</i> nuclear polyhedrosis virus	0.00	0.00	0.00	0.00	1.54	0.00	3.08	1.54	0.00	3.08	0.00	40.00	16.92	10.77	1.54	0.00	10.77;	
3.08	0.00	0.00	0.00	0.00	0.00	0.00	1.54	0.00	3.08	1.54	0.00	3.08	0.00	40.00	16.92	10.77	1.54	0.00	10.77;	

---

## References

1. Mayoral, E.; Dubchak, I.; Muchnik, I. *Ismb* 1995, 3, 240.
2. Muskal, S. M.; Kim, S. H. *J Mol Biol* 1992, 225, 713.
3. Lesk, A. M.; Boswell, D. R. *Bioessays* 1992, 14, 407.
4. Tuckwell, D. S.; Humphries, M. J.; Brass, A. *Comput Appl Biosci* 1995, 11, 627.
5. Garnier, J.; Gibrat, J. F.; Robson, B. *Methods Enzymol* 1996, 266, 540.
6. Fasman, G. D., Ed. *Prediction of Protein Structure and the Principles of Protein Conformation*; Plenum Press: New York, 1989.
7. Jaroszewski, L.; Rychlewski, L.; Zhang, B.; Godzik, A. *Protein Sci* 1998, 7, 1431.
8. Fischer, D.; Eisenberg, D. *Protein Sci* 1996, 5, 947.
9. Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. *Annu Rev Phys Chem* 1997, 48, 545.
10. Nakashima, H.; Nishikawa, K.; Ooi, T. *J Biochem (Tokyo)* 1986, 99, 153.
11. Hatch, F. T. *Nature* 1965, 206, 777.
12. Harding, J. J. *Biochem J* 1984, 217, 339.
13. Chou, P. Y. In *Prediction of Protein Structure and the Principles of Protein Conformation*; Fasman, G. D., Ed.; Plenum Press: New York, 1989, p. 549.
14. Zhang, C. T.; Chou, K. C. *Protein Sci* 1992, 1, 401.
15. Dubchak, I.; Holbrook, S. R.; Kim, S. H. *Proteins* 1993, 16, 79.
16. Nakai, K.; Kidera, A.; Kanehisa, M. *Protein Eng* 1988, 2, 93.
17. Chun-Ting, Z.; Xinhua, X.; Genfa, Z. *Eur J Biochem* 1992, 210, 747.
18. Landes, C.; Risler, J. L. *Comput Appl Biosci* 1994, 10, 453.
19. Chou, K. C. *Proteins* 1995, 21, 319.
20. Chou, K. C. *FEBS Lett* 1995, 363, 127.
21. Chou, K. C.; Zhang, C. T. *Crit Rev Biochem Mol Biol* 1995, 30, 275.
22. Chou, K. C.; Zhang, C. T. *J Biol Chem* 1994, 269, 22014.
23. Chou, K. C.; Maggiora, G. M. *Protein Eng* 1998, 11, 523.
24. Sammon, J. W. *IEEE Trans Comp* 1969, C-18, 401.
25. Agrafiotis, D. K. *Protein Sci* 1997, 6, 287.
26. Barlow, T. W.; Richards, W. G. *J Mol Graph* 1995, 13, 373.
27. Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. *J Mol Biol* 1995, 247, 536.
28. Cornish-Bowden, A. *Biochem J* 1983, 213, 271.
29. Szpankowski, W. *SIAM J Comput* 1993, 22, 1176.